学校编码：10384　　分类号_____密级_____
学　　号：22320051302503　　UDC _____

厦门大学

硕 士 学 位 论 文

基于中日韩的多国语言编码系统的研究与实现

The Research and Implementation of multi-language system based on Chinese, Japanese, and Korean

黄志勇

指导教师姓名：吴顺祥 教授

专 业 名 称：系 统 工 程

论文提交日期：2008 年 5 月

论文答辩时间：2008 年 5 月

学位授予日期：2008 年 月

答辩委员会主席：_____

评　　阅　　人：_____

2008 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

目前计算机取证的重点主要集中在简体中文操作系统中。随着国际交流的加强和我国改革开放的深入，以及涉及外国人的计算机案件量的不断增加。各省各地市缺乏精确高效的多语言后处理手段，使得涉及多语言的电子数据后处理工作诸多不便。当遇到此类案件的时候，往往让调查人员花费大量的时间。同时由于各自使用的手段方法不统一，也难于将各自的应用技术或方法推广与普及。

本文研究的系统正是针对上面的一些问题，指出国内的信息分析系统和取证系统的不足，对中、日、韩三国的计算机常用编码规则进行深入分析与研究。通过给出取证搜索关键字在计算机中的可能的 16 进制编码的方式，让调查人员可以利用标准的计算机取证软件对 16 进制编码进行关键字搜索，从而实现案件线索的快速定位。为取证工作简化工作模式，在提高现有的工作质量前提下提高工作效率。

本文首先阐述了基于中日韩的多国语言编码系统的重要意义，介绍了国内外计算机取证调查分析的现状；接着，详细研究了中日韩三国所涉及到的各种编码；然后，对这些编码的内容、特点、编码/解码原理及方法进行了深入的分析：本文以 Unicode 为核心编码，深入研究了中文的 GB2312、BIG5、GBK 等，日文的 JIS、SHIFT_JIS、EUC-JP 等，韩文的 EUC-KR 等本地编码规则，以及写字板的 rtf，邮件的 base64 与 QP 编码，UTF-8 等二次编码规则和 Unicode 编码，本系统支持直接批量导出为重要取证软件 EnCase 的关键字搜索脚本；然后，介绍了系统设计、运行环境、和具体实现的基本情况等；最后，对本系统的研究开发情况进行了总结，并就开发中的不足之处提出了下一步的努力方向。

经过测试，本文研究的基于中日韩的多国语言编码系统能有效的提高取证人员的工作效率，扩大取证搜索的范围，为有力打击计算机犯罪活动提供了很好的应用工具。本系统还可用于纠正乱码，增强国际信息交流等多方面。同时该系统得到了很好的推广应用，具有较高的社会意义和实用价值。

关键词：编码转换；解码；计算机取证；Unicode；EnCase

ABSTRACT

At present, the computer forensics put it's emphases on Simplified Chinese in windows. Along frequently international communication and embedded reform and open policies, the cases about foreigners are increasing continuously. Because of the lack of precise multiplicate language tool to deal with, the investigators have to spend much time on the disposal of computer data. At the same time, to use the different methods make the technics difficult to spread.

For the questions above, this paper analysed the shortage of information analysis and computer forensics system, and researched the computer code regulations in China, Japan, and Korea. Through giving the hex code of possible key words for search, let the investigators use standard software for computer forensics to search the key words easily and fleetly. So, it can predigest the work of forensics, and improve the work efficiency under the premise of work quality.

This paper described the significance of the multi-language system based on Chinese, Japanese, and Korean firstly, introduced the island and overseas status quo of computer forensics. Then, it traversed most kind of codes which the computers in China, Japan, and Korea refer to. Then, used Unicode as kerneled code, lucubrated to the local codes, such as GB2312, BIG5, GBK, etc. in Chinese, JIS, SHIFT_JIS, EUC-JP, etc. in Japanese, EUC-KR, etc. in Korean; the quadratic codes such as rtf document, base64 and QP code; and the Unicode. The system sustain to export a batch EnCase' scrips of key words search. Then the paper introduced the design of system, the environment to run, and basic circs to implement concretely. Finally, summarized the system development and on the lack propose the next step to make efforts.

Through testings, the system discussed in paper can effectively improve the work efficiency, enlarge the range of forencis search, and provide a good tool for striking computer crimes. It also can correct unreadable codes, enhance the communication of international information. It has been generalized widely, and has high significance for society, and applied value.

Keywords: code transition; decode; computer forencis; Uncode; EnCase

目　录

Contents

第一章 绪 论

1.1 研究背景

根据美国联邦调查局的估算，通过计算机、通信设备、互联网实施的高科技犯罪罪每年至少给美国社会造成 3000 亿美元的损失。在全球范围内，此类日益增长的高科技犯罪已经成为各国普遍面临的现实问题。

作为打击高科技犯罪的主要手段之一，电子物证鉴定、数据恢复、电子邮件分析、手机取证分析、互联网在线取证调查等计算机法证技术也越来越受到关注。计算机法证技术就是通过一定的方法和程序，对各种数据源以及各种形式的电子数据进行固定、分析、鉴定和出示，揭示并再现利用计算机或互联网进行犯罪的行为和过程，并发现各种潜在威胁[17,33,40]。

根据业内专家透露，目前国内对于计算机法证技术的应用需求相当迫切，尤其是在计算机犯罪、数字版权保护等方面。电影《满城尽带黄金甲》的制作方就曾经在电影公映期间，委托国内相关企业采用计算机法证技术进行网络维权。搜索查找电影上映期间非法提供没有授权的影片下载网站，并采集相关的电子证据，以此来警示对方，或是作为最终采取法律措施的证据。

由于我国基于计算机法证技术的研究还处于起步阶段，整体技术水平与国外相比还有一定的差距，尤其是在关于计算机法证技术的立法方面。有专家指出，立法的滞后已经使计算机法证技术在证据原始性、软件同源性、信息彻底性、数据关联性和司法规范性等方面受到了挑战。

国内某知名网络游戏公司就曾经因为证据原始性问题在与“私服”对簿公堂中败诉，败诉的原因就是由于“私服”查处封存的时间和之后采集电子数据的时间不统一，法庭无法判定证据的原始性，从而裁定“私服”一方胜诉。

从 1997 年就开始高科技犯罪案件调查工作的香港警务处商业犯罪调查科总督察 Paul Jackson 表示，对于计算机法证技术最重要的一条原则就是原始数据不可改变。这其中包含着 3A 原则：获取（Acquire）原始证据不能有更改和损坏；鉴定（Authenticate）所恢复的数据必须与原始数据一致；分析（Analyze）数据过程中不能对其进行修改。只有做到这三条原则，才能使计算机法证技术实现取证的终极目标。在实际操作中，除了要有相应的技术保障外，针对计算机法证技术的相关标准和立法也是取保上述原则得

到执行的重要因素[18-20]。

1.2 研究现状

在国际上，军队、警察、海关、反贪、金融、税务、律师、保险等部门是电子证据的主要应用部门。而与此同时，由于各个行业涉及计算机、局域网、互联网的高科技犯罪、商业欺诈、白领犯罪等行为越来越多，因此有越来越多的咨询顾问公司、商业调查机构、会计师行、私人调查公司开始从事电子证据服务。

目前计算机取证的重点主要集中在简体中文操作系统中。随着国际交流的加强和我国改革开放的深入，以及涉及外国人的计算机案件量的不断增加。目前，在各省各地市缺乏精确高效的多语言后处理手段。为涉及多语言的电子数据后处理工作带来诸多不便。当遇到此类案件的时候，往往不仅让调查人员花费大量的时间。同时由于各自使用的手段方法不统一也难于将各自的应用技术或方法推广与普及。

因此，其他语言也将是调查取证的重要对象。尤其是中国的邻国——韩国，日本等。所以我国（尤其是沿海地区）急需一套简单易用的多语言编码系统来辅助调查取证工作的开展。但限于目前国内涉及外国人的计算机案件基本仅仅只有国家安全部门参与，因此该软件的应用对象比较具有针对性。但该软件的应用在安全系统内部可以进一步引申到国际互连网网络信息的监控。

1.2.1 计算机取证

从技术角度看，计算机取证是分析硬盘、光盘、软盘、Zip 磁盘、U 盘、内存缓冲和其他形式的储存介质以发现犯罪证据的过程，即计算机取证包括了对以磁介质编码信息方式存储的计算机证据的保护、确认、提取和归档。取证的方法通常是使用软件和工具，按照一些预先定义的程序，全面地检查计算机系统，以提取和保护有关计算机犯罪的证据。

计算机取证主要是围绕电子证据进行的。电子证据也称为计算机证据，是指在计算机或计算机系统运行过程中产生的，以其记录的内容来证明案件事实的电磁记录。多媒体技术的发展，电子证据综合了文本、图形、图像、动画、音频及视频等多种类型的信息。与传统证据一样，电子证据必须是可信、准确、完整、符合法律法规的，是法庭所能够接受的。同时，电子证据与传统证据不同，具有高科技性、无形性和易破坏性等特